I. INTRODUCTION

A hearing aid is intended to improve speech intelligibility for the hearing-impaired user, but it is often difficult to predict the amount of benefit that will be obtained in practice. Intelligibility is reduced not only by the hearing loss, but also by ambient noise, room reverberation, and the noise and nonlinear distortion introduced by the hearing aid. The desire to assess speech intelligibility under adverse conditions has led to the development of several measurement and prediction procedures. Some of the problems affecting speech intelligibility, such as additive noise and bandwidth reduction, have been addressed by the speech intelligibility index (SII) (ANSI S3.5-1997), and reverberation has been addressed by the speech transmission index (STI) (Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985). However, there is no metric that can successfully deal both with noise and with the various types of nonlinear distortion that can occur in a hearing aid or communication system.

An intelligibility-estimation procedure for hearing aids must satisfy several criteria. First, the metric must be applicable to systems, like hearing aids, that have frequency-dependent magnitude and phase transfer functions. A flat frequency response, often assumed in telephony and for digital speech coding systems, cannot be assumed for a hearing aid. Second, in the presence of additive noise without distortion, the metric should yield intelligibility predictions that are consistent with those produced by the SII calculation (ANSI S3.5-1997). Third, the method should be applicable to speech as the test signal; much of the nonlinear signal processing in hearing aids is specifically designed for speech as the input and will respond differently to other types of excitation such as pure tones, multitone complexes, or broadband noise.

The goal of the work reported in this paper is to derive a speech intelligibility metric that generalizes the SII. The specific objectives of the work are twofold: first, the metric should accurately predict the effects of broadband noise and distortion for speech reproduced by hearing aids and other communication systems, and, second, the metric should be accurate for hearing-impaired as well as normal-hearing listeners.

A. Speech intelligibility in normal-hearing listeners

There are several procedures for estimating speech intelligibility in normal-hearing listeners. The best known are the articulation index (AI) (French and Steinberg, 1947; Kryter, 1962), which has been further developed to produce the speech intelligibility index (ANSI S3.5-1997), and the speech transmission index (Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985). The SII measures the signal-to-noise ratio (SNR) on a dB scale in each frequency band. The calculation procedure then adjusts for auditory threshold and for frequency-domain masking effects and sums the weighted SNR across frequency to produce the intelligibility estimate. The STI uses amplitude-modulated speech-shaped noise as the excitation. Its calculation procedure measures the modulation depth in each frequency band in the listening environment and forms a weighted sum across frequency to give an estimate of the SNR. The STI calculation can also be performed using speech as the stimulus (Payton and Braida, 1999). The SII is effective in stationary additive noise and for filtering that restricts the speech bandwidth, and the STI is also effective in reverberation.

The SII procedure embodies a simplified model of the auditory periphery. The critical band formulation of the SII
procedure given in Table I of the standard (ANSI S3.5-1997) uses the critical bands of Zwicker (1961), which give a representation of auditory frequency analysis. The auditory threshold for normal and impaired hearing is incorporated into the calculation in Eq. (10) of the standard, and upward spread of masking for normal ears is included in Eqs. (6)–(9) of the standard. Because of this implicit auditory model, the SII procedure can be extended to include conditions not incorporated into the standard as long as the basic physical and perceptual assumptions are not violated.

Examples of previous extensions to the SII are the work of Kates (1987) and Rhebergen and Versfeld (2004). To compute the SII, the speech power spectrum and the noise power spectrum are averaged across the entire duration of the stimulus, and the SII is then calculated using these long-term averages. The SII computed this way is accurate for stationary noise, but is inaccurate for fluctuating noise. Kates (1987) evaluated an adaptive noise-cancellation system by dividing the speech into segments, computing the SII for each segment, and then plotting the change in the SII as a function of time to illustrate the system convergence. Rhebergen and Versfeld (2004) have extended the SII approach to predict the effects of fluctuating noise. They propose dividing the noisy speech into segments, computing the SII for each segment, and then averaging the SII values. The SII assumptions concerning auditory frequency resolution, threshold, and masking are preserved in their procedure, but the analysis adds additional understanding of the effects of signal fluctuations and the ability to listen in the noise valleys.

The principles underlying the SII have also been applied to estimating speech intelligibility and quality for distortion mechanisms such as symmetric peak clipping. Steeneken and Houtgast (1980), for example, show that the modulation transfer function can accurately predict speech intelligibility for peak-clipped speech. Kates and Kozma-Spytek (1994) studied speech subjected to low-frequency or high-frequency amplification followed by symmetric peak clipping. They computed the signal-to-distortion ratio (SDR) from the magnitude-squared coherence (MSC), and then derived a speech-quality metric based on a weighted sum of the SDR across frequency. This metric, which is similar to the frequency-dependent weighting of the SNR used in the SII calculation, was accurate in predicting speech quality for peak clipping even though auditory masking was not included. An extension of this approach (Kozma-Spytek et al., 1996) was accurate in predicting speech quality for peak-clipped speech for hearing-impaired listeners, although some of the hearing-impaired subjects appeared to concentrate more on the low frequencies and other subjects appeared to concentrate more on the high frequencies in making their judgments.

B. Speech intelligibility in hearing-impaired listeners

Speech intelligibility in hearing-impaired listeners has been studied for conditions of bandwidth reduction and additive noise, but only rarely for distortion. For additive noise and bandwidth reduction, the SII (ANSI S3.5-1997) procedure has been found in general to overestimate speech intelligibility for hearing-impaired listeners (Pavlovic, 1984; Ludvigsen, 1987; Hornsby and Ricketts, 2003). One solution is to apply a desensitization factor to the SII calculation, reducing the estimated intelligibility in proportion to the measured hearing loss (Pavlovic et al., 1986; Magnusson et al., 2001). The desensitization factor does not explain the reason for the discrepancy between the SII and the observed intelligibility, but it does provide an accurate empirical correction. The desensitization factor accuracy can be further improved by applying individually determined frequency-dependent proficiency factors in the SII calculation (Ching et al., 1998).

A second solution to the problem of overestimating the intelligibility for hearing-impaired listeners is to modify the effective signal-to-noise ratio (SNR) used in the SII calculation. For example, the upward spread of masking function used in the SII calculation is based on averaged measurements of normal-hearing subjects. Increasing the spread of masking for the impaired ear reduces the effective signal-to-noise ratio (SNR) above the frequency region of the masker. The reduction in SNR reduces the calculated SII value. Rankovic (1998) used individual frequency-domain masking functions in applying the Fletcher and Galt (1950) calculation procedure to hearing-impaired listeners and found that the individualized calculations improved the accuracy of the intelligibility predictions without needing a desensitization factor. Hornsby and Ricketts (2003) measured pure-tone detection thresholds in speech-shaped noise for each hearing-impaired listener and used these data to increase the effective masking spectrum in the SII calculation. This modification of the SII implicitly includes the increase in spread of masking in the impaired ear and thus reduces the effective SNR. The modification resulted in intelligibility predictions for hearing-impaired subjects, having flat losses, that were comparable in accuracy to those for normal-hearing listeners, again without the need for a desensitization factor.

Crain and Van Tasell (1994) studied speech intelligibility in hearing-impaired subjects for peak-clipping distortion. They found that peak-clipping distortion increased the speech reception threshold (SRT) more for hearing-impaired listeners than for normal-hearing listeners. Crain and Van Tasell compared the SRT for no distortion to the SRT for infinite peak clipping after equating the levels of presentation. They found that the SRT for the hearing-impaired subjects shifted by 5 dB while that for the normal-hearing subjects shifted by only 2 dB. The shift in SRT was also found to be highly correlated with the reduction in the coherence between the input and distorted signals.

C. Test signal

Speech is the desired test signal for most types of hearing-aid processing. In many nonlinear processing systems, such as the noise suppression and multiband compression found in digital hearing aids, the signal processing behavior depends on the characteristics of the signal. For example, replacing speech with the amplitude-modulated speech-shaped noise used in the STI measurement (Steeneken and Houtgast, 1980) or with a synthesized approximation to speech (CCITT, 1986) will result in different frequency- and time-dependent gains than would occur for...
real speech. Thus estimating the distortion that will occur for speech in a nonlinear speech-processing system requires the use of speech or, equivalently, a speechlike test signal that has statistics very close to those of real speech.

### D. Coherence measurements and hearing aids

Coherence measurements are becoming important in evaluating hearing aids. Coherence measurement procedures using a broadband input signal have been developed for hearing aids (Preves et al., 1989; Dyrland, 1989; Preves, 1990; Kates, 1992; Fortune and Preves, 1992; Schneider and Jamieon, 1995; Kates, 2000), and measurement procedures were standardized in ANSI S3.42-1992. The use of a broadband stimulus is essential for determining the response of multi-channel compression hearing aids; in these systems the gain as a function of frequency depends on both the signal amplitude and spectral distribution. Coherence is attractive for determining the nonlinear behavior of such systems because it measures all forms of distortion and not just the harmonic distortion traditionally measured (ANSI S3.22-1996).

An implicit assumption in using coherence to measure the nonlinear system behavior is that equal reductions in coherence will cause equal reductions in intelligibility or sound quality. In dealing with a single distortion mechanism, such as amplifier saturation, a correlation between a reduction in sound quality and a reduction in coherence is often found (Palmer et al., 1995). However, different distortion mechanisms may reduce the MSC by the same amount yet have different perceptual consequences. The relationship between coherence and intelligibility and sound quality has not been investigated for different forms of distortion, and the assumption that equal reductions in MSC are perceptually equivalent has not been tested.

### E. Summary of the paper

In this paper the SII approach is extended to estimate intelligibility for peak clipping and center clipping distortion as well as for additive noise. The new procedure is a modified form of the SII procedure wherein the standard speech signal-to-noise ratio estimate is replaced by the signal-to-distortion ratio computed from the MSC. The MSC is used because it is immune to linear transformations of the signal, is reduced by nonlinear distortion as well as by additive noise, and because it is applicable to any excitation signal, including speech. The SII is used as the basis of the metric because of its established accuracy in predicting intelligibility for additive noise and because it includes a model of auditory masking as well as the threshold shift in impaired ears. The new metric is tested with both normal-hearing and hearing-impaired listeners.

The remainder of this paper describes the evaluation and modeling of speech intelligibility scores from normal-hearing and hearing-impaired individuals. The stimuli used in the evaluations are sentences subjected to different forms of noise and distortion. The first experiment uses normal-hearing listeners. The subject intelligibility scores are modeled using a modification of the SII procedure in which coherence is used to estimate the effects of the noise and distortion on the speech. It is shown that the predictions using the coherence SII do not adequately model the intelligibility results. A further modification, based on computing the coherence SII separately for low-, medium-, and high-level segments of each sentence, is shown to be much more accurate in predicting the intelligibility scores.

The second experiment uses hearing-impaired listeners. The processed sentence stimuli are the same as used for the normal-hearing subjects. However, the hearing-impaired subjects are provided with amplification to compensate for each individual hearing loss. The intelligibility scores for the hearing-impaired subjects are predicted using the three-level coherence SII procedure with the same weights for the low-, mid-, and high-level segments of the sentences as used for the normal-hearing subjects. The only modification is the inclusion of the individual hearing loss in the coherence SII calculations. The new procedure is shown to be just as accurate in predicting the intelligibility for the hearing-impaired listeners as it is for the normal-hearing listeners.

### II. EXPERIMENT I: NORMAL-HEARING LISTENERS

#### A. Listeners

Thirteen adult listeners with normal hearing (age range 18 to 45 years) participated. All listeners had thresholds of 20 dB HL (ANSI, 1989) or better at octave frequencies from 250 to 8000 Hz, inclusive. Listeners were tested individually in a double-walled sound booth and were compensated $10/hour for their participation.

#### B. Test materials

The test materials consisted of the Hearing-in-Noise-Test (HINT) sentences (Nilsson et al., 1994). The sentences were digitized at a 44.1-kHz sampling rate and down-sampled to 22.05 kHz to reduce computation time. There are 25 lists with ten sentences per list, giving a total of 250 sentences. Each test sentence was combined with additive noise or was subjected to symmetric peak-clipping distortion or symmetric center-clipping distortion. The two forms of clipping were chosen as examples of memoryless nonlinearities that have been used in previous investigations (Licklider, 1946). Peak clipping is related to arithmetic, amplifier, and receiver saturation in a hearing aid. Center clipping is related to noise-suppression systems that reduce the amplitude of low-level portions of the signal.

The additive noise was extracted from the opposite channel of the HINT test compact disc. The noise has the same long-term spectrum as the sentences. The signal-to-noise ratio (SNR) was determined by computing the root-mean-squared (RMS) power of each sentence, ignoring the silent intervals at the beginning and end of the sentence, and adjusting the noise power over the same interval to give the desired SNR. The combined signal-plus-noise power over the duration of each sentence was then adjusted to give a presentation level of 65 dB SPL.

The peak-clipping and center-clipping distortion thresholds were set using the histogram of the magnitudes of the signal samples for each sentence. The silent intervals at the
beginning and end of each sentence were discarded, and the cumulative distribution of the absolute values of the signal samples was then computed for the sentence. The clipping threshold was set as a percent of the cumulative magnitude histogram for the sentence. For symmetric peak clipping, the clipping operation is given by

\[ y(n) = \begin{cases} 
  c, & x(n) > c, \\
  x(n), & -c \leq x(n) \leq c, \\
  -c, & x(n) < -c 
\end{cases} \]

where \( x(n) \) is the speech input, \( y(n) \) is the distorted output, and \( c \) is the clipping threshold. The symmetric center clipping operation is given by

\[ y(n) = \begin{cases} 
  x(n), & x(n) > c, \\
  0, & -c \leq x(n) \leq c, \\
  x(n), & x(n) < -c 
\end{cases} \]

The distorted sentences were then readjusted to give an average level of 65 dB SPL.

For each individual listener, 24 lists with ten sentences per list were selected for processing, with eight lists randomly selected for each of the three distortion conditions. Eight levels of distortion were evaluated for each distortion type and the distortion levels were randomized within each block. The distortion levels were peak clipping with the clipping threshold set to \(0\%, 50\%, 75\%, 90\%, 95\%, 98\%, 99\%, 100\%\) of the cumulative magnitude histogram of each sentence; center clipping with the clipping threshold set to \(0\%, 50\%, 70\%, 80\%, 85\%, 90\%, 95\%, 98\%\) of the cumulative magnitude histogram; and additive noise with the SNR set to \(-5, 0, 5, 10, 15, 20, 30, 100\) dB. Note that the effects of peak clipping are reduced as the clipping threshold is increased, while the effects of center clipping are increased as the clipping threshold is increased.

C. Procedure

Listeners participated in 1-h sessions. During the session, the listeners were presented with three blocks of 80 sentences, with one block in each of the three distortion conditions. The order of presentation of the three distortion conditions was randomized across listeners and across sessions. The order of presentation of sentences within a block was also randomized for each listener. During the sessions, listeners verbally repeated each sentence after it was presented. The tester then scored the proportion of complete HINT sentences that were correctly repeated by the listener.

For listener presentation, the digitally stored speech stimuli went through a digital-to-analog converter (Tucker-Davis Technologies RP2), an attenuator (TDTS PA5), and a headphone buffer amplifier (TDTS HB7). Finally, the stimuli were presented monaurally to the left ear of each listener through a TDH-49 earphone. All stimuli were presented to the listeners at an equalized-RMS level of 65 dB SPL.

D. Intelligibility scores

The intelligibility scores for the normal-hearing subjects are plotted in Fig. 1. The scores for each listener were first averaged across the ten sentences used for each combination of distortion type and level. The average intelligibility scores for each condition were then averaged across the 13 listeners. The average intelligibility (proportion sentences correct) is plotted as a function of SNR in dB (right panel) and as a function of the clipping threshold expressed as a percent of the cumulative amplitude distribution (left panel). The error bars show the standard deviation across the 13 subjects.

For the additive noise, the intelligibility increases as the SNR increases. The general shape of the curve is consistent with previous intelligibility results for the HINT sentences in additive speech-shaped noise (Nilsson et al., 1994; Eisenberg et al., 1998). The speech intelligibility is a steep function of SNR over the range of \(-5\) to \(+5\) dB; the intelligibility goes from 22% to 97%, giving an average slope of
The intelligibility reaches an asymptote of nearly perfect sentence recognition above the 5-dB SNR.

The worst case intelligibility for peak clipping is 72% correct, obtained for infinite clipping (replacing the speech signal samples by their signs). The intelligibility improves as the clipping threshold increases. The high intelligibility for peak clipping indicates that much of the vowel and consonant structure of the speech is preserved even for infinite clipping. For center clipping, the intelligibility remains high for a clipping threshold up to 70%, and then decreases rapidly with increasing clipping threshold. At a clipping threshold of 98% all but the highest peaks of the speech signal have been removed, resulting in an intelligibility of 0%. These results are consistent with previous results for peak-clipped and center-clipped speech (Licklider, 1946).

III. COHERENCE

The coherence function (Carter et al., 1973) is the normalized cross-spectral density given by

$$\gamma(\omega) = \frac{S_{xy}(\omega)}{|S_x(\omega)S_y(\omega)|}$$

where $S_{xy}(\omega)$ is the cross-spectral density between the two stationary random processes $x(n)$ and $y(n)$ having autospectral densities $S_x(\omega)$ and $S_y(\omega)$, respectively. The magnitude-squared coherence function (MSC) is then given by

$$|\gamma(\omega)|^2 = \frac{|S_{xy}(\omega)|^2}{|S_x(\omega)|^2|S_y(\omega)|^2}$$

with $0 \leq |\gamma(\omega)|^2 \leq 1$.

When $x(n)$ is the input to a system and $y(n)$ is the output, the MSC represents the fraction of the output signal power that is linearly dependent on the input at each analysis frequency. The complementary fraction $1 - |\gamma(\omega)|^2$ gives the output power at each frequency that is unrelated to the input at that frequency and thus represents the nonlinear distortion and noise. The signal-to-noise and distortion ratio (SDR) is therefore given by (Carter et al., 1973)

$$\text{SDR}(\omega) = |\gamma(\omega)|^2(1 - |\gamma(\omega)|^2).$$

When $|\gamma(\omega)|^2 = 1$, all of the output is linearly related to the input and the SDR approaches infinity. When $|\gamma(\omega)|^2 = 0$, there is no relationship between the output and the input and the SDR is zero as well.

The MSC is typically estimated using fast Fourier transform (FFT) techniques (Carter et al., 1973). The data sequences $x(n)$ and $y(n)$ are divided into a number $M$ of overlapping windowed data segments. The cross-spectrum and autospectra are computed for each segment using the FFT and are then averaged across segments, and the MSC is computed from the averages. For $M$ data segments, the estimated MSC is given by

$$|\gamma(k)|^2 = \frac{\sum_{m=0}^{M-1} X_m(k)Y_m^*(k)^2}{\sum_{m=0}^{M-1}|X_m(k)|^2 \sum_{m=0}^{M-1}|Y_m(k)|^2},$$

where the asterisk denotes the complex conjugate. $X_m(k)$ and $Y_m(k)$ are the spectra of the $m$th windowed data segments of $x(n)$ and $y(n)$, respectively, computed using the FFT algorithm where $k$ is the FFT bin index. Typically, a Hann or Hamming data window and 50% overlap are used for the data segments. The estimated MSC is subject to bias effects due to the finite number of segments used in the FFT procedure and to the interaction of the processing system delay with the segment length. For the greatest accuracy, unbiasing techniques that temporally align the input and output sequences should be used when computing the MSC (Kates, 1992).

The coherence is unaffected by a linear transformation of the signal. A linear filter will increase the gain at each frequency, but the change in gain affects the numerator and denominator of Eq. (6) by equal amounts. Nonlinear transformations, however, will reduce the MSC (Kates, 2000). Thus the MSC is reduced by additive noise, but it is also reduced by nonlinear distortion such as peak clipping, center clipping, and by dynamic-range compression.

IV. THE COHERENCE SII

A. Calculation procedure

The SII standard (ANSI S3.5-1997) allows the SNR to be calculated using the speech and noise spectra measured in octave, one-third octave, or auditory critical bands. The critical-band SII procedure is used in this study, and the filter center frequencies and bandwidths are given in Table I of ANSI S3.5-1997. The critical band filter shapes used in this study are the simplified ro-ex filters suggested by Moore and Glasberg (1983). Let $q_j$ be the center frequency in kHz of the $j$th critical band, with the band center frequency shifted to match the center frequency of the closest FFT analysis bin. The filter bandwidth from the ANSI Table I is denoted by $b_j$; the bandwidths correspond to the classical critical bands (Zwicker and Terhardt, 1980). The ro-ex filter parameter is given by

$$p_j = \frac{4(1000q_j)}{b_j},$$

where the factor of 1000 converts the filter center frequency to Hz. The simplified ro-ex filter is then given by

$$W_j(f) = (1 + p_j g) \exp(-p_j g),$$

where $g$ is the normalized deviation from the filter center frequency given by

$$g = |1 - f|q_j|$$

and $f$ is the FFT bin center frequency in kHz.

The standard SII calculation assumes that the speech and noise spectra have been measured separately. Let $P(k)$ be the speech power spectrum and $N(k)$ be noise power spectrum computed separately using FFT techniques and where $k$ is the FFT bin index. The SNR for the $j$th critical band is then given by

$$\text{SNR}(j) = \sum_{k=0}^{K} W_j(k)P(k)$$

where $W_j(k)$ is the frequency response of the ro-ex filter. In a system that includes distortion, the noise-plus-distortion spectrum is not directly available and must therefore be estimated from the MSC. The speech power spectrum computed using the MSC is given by

$$P(k) = \sum_{k=0}^{K} W_j(k)N(k).$$
and the noise power spectrum is given by
\[ \hat{N}(k) = |\gamma(k)|^2 S_{yy}(k), \]  
where \( S_{yy}(k) \) is the output power spectral density estimated using the FFT. The SDR estimated using the MSC is thus
\[ \text{SDR}(j) = \frac{\sum_{k=0}^{K} |W_j(k)\gamma(k)|^2 S_{yy}(k)}{\sum_{k=0}^{K} |W_j(k)(1 - |\gamma(k)|^2)S_{yy}(k)|}. \]  

The calculation of Eq. (13) uses the output power spectrum and the coherence, and thus requires access to the input and output signals of the system. However, the system under test can be treated as a “black box;” the procedure does not require knowledge of the distortion mechanism or separate measurements of the noise or distortion signals. The standard SII is computed using the procedures of ANSI S3.5-1997 and the SNR estimate of Eq. (10). The coherence-based SII is computed using the same procedures, but with the speech power spectrum replaced by Eq. (11), the noise power spectrum replaced by Eq. (12), and the SNR replaced by the SDR estimate of Eq. (13).

B. Validation for additive noise

The SDR computed using Eq. (13) and the SNR computed using Eq. (10) should agree for additive noise. That is, the SDR should give the same results as the SNR if only additive noise (and no distortion) is present in the system being measured. If the SDR equals the SNR, the standard and coherence approaches for computing the SII will yield identical results for additive noise.

The SDR and SNR are compared for additive low-pass filtered noise in Fig. 2. The input was a concatenation of the ten sentences in the HINT test list 1 (Nilsson et al., 1994) read by a male talker. The gaps between the sentences were removed. The sentences were digitized at a 44.1-kHz sampling rate and down-sampled to 22.05 kHz. The stimulus duration was approximately 18 s. Figure 2 is for the concatenated sentences with low-pass filtered (one-pole filter at 900 Hz) noise at a 10-dB SNR. The spectra used a FFT size of 32 ms, which is 706 samples at the 22.05-kHz sampling rate. The FFTs were computed using 706-point segments having 50% overlap and weighted with a Hamming window. A new FFT was thus computed every 16 ms, and the FFT had a frequency resolution of 31.23 Hz.

The SDR and SNR are nearly identical, with the largest difference between the two being less than 0.5 dB. Thus the requirement that the SDR and SNR agree for additive noise is satisfied. The SII values computed using the SDR and SNR are identical, being 0.671 for a signal level of 65 dB SPL.

C. Intelligibility predictions

The proportion of the HINT sentences identified correctly is plotted in Fig. 3 as a function of the coherence SII (CSII) for the noise and distortion conditions. The results for the additive speech-shaped noise are in reasonable agreement with the results of Eisenberg et al. (1998); an SII of 0.5 corresponds to approximately 90% sentences correct for both sets of data. A CSII of 0.2 corresponds to approximately 35% correct in the results plotted in Fig. 3, while Eisenberg et al. (1998) found that an SII of 0.3 gave approximately the same performance.

The intelligibility for peak-clipped speech is similar to that for the additive noise for the same CSII values. The worst case for peak clipping is infinite clipping, which results in a CSII value of approximately 0.3. But even for this worst case there does not appear to be a significant difference between the additive noise and the peak-clipping intelligibility results. For center clipping, however, the intelligibility is significantly worse than for additive noise or peak clipping at the same CSII values. A possible explanation is that center clipping removes a large fraction of the consonant information in the speech signal, but the CSII calculation is dominated by the more-intense portions of the signal and thus primarily reflects modifications of the vowels. Thus the assumption that equal CSII values will give equal intelligibility.

FIG. 2. SDR (solid line), SNR (dashed line), and the difference between them (dot-dashed line) as a function of the frequency band number for the concatenated HINT list 1 sentences. Additive noise low-pass filtered at 900 Hz is present at an SNR of 10 dB.

\[ \hat{P}(k) = |\gamma(k)|^2 S_{yy}(k) \]  

and the noise power spectrum is given by
\[ \hat{N}(k) = [1 - |\gamma(k)|^2]S_{yy}(k), \]  
where \( S_{yy}(k) \) is the output power spectral density estimated using the FFT. The SDR estimated using the MSC is thus
\[ \text{SDR}(j) = \frac{\sum_{k=0}^{K} W_j(k)|\gamma(k)|^2 S_{yy}(k)}{\sum_{k=0}^{K} W_j(k)(1 - |\gamma(k)|^2)S_{yy}(k)}. \]  

The proportion of the HINT sentences identified correctly plotted versus the full-level coherence SII calculation for normal-hearing subjects. The triangles, circles, and squares show the average results for the additive noise, peak-clipping, and the center clipping, respectively.

FIG. 3. Proportion of the HINT sentences identified correctly plotted versus the full-level coherence SII calculation for normal-hearing subjects. The triangles, circles, and squares show the average results for the additive noise, peak-clipping, and the center clipping, respectively.
A. Coherence SII calculation

V. THREE-LEVEL CSII

The goal of the new speech intelligibility metric is to derive a single model that accurately predicts the effects of the additive noise, peak clipping, and center clipping used in the experiment. The coherence SII, as implemented above, is inadequate in representing the effects of all three noise and distortion conditions. More information about how both noise and distortion affect the speech signal is needed to predict the subject intelligibility results. This additional information can be provided by dividing the speech signal envelope into three amplitude regions and performing a separate CSI calculation for each region. The three CSI values are then combined to model the intelligibility scores.

A similar multi-level procedure was proposed for estimating perceived speech intelligibility for voice communication systems (CCITT, 1986). In the CCITT (1986) procedure, a synthetic speech signal is divided into overlapping 32-ms windowed segments. The mean-squared level of each segment is computed, and a histogram constructed from the segment magnitudes. Each segment is then assigned to one of four quartiles of the histogram. For the segments in each quartile, the MSC is computed and used to scale the output power spectrum for the quartile to give the speech and noise power spectral estimates. The noise power spectrum is combined with the auditory threshold to give a masking function, the speech power spectrum is converted to sensation levels, and an intermediate score is computed in a manner similar to the SII. The intermediate scores in each quartile are then summed using a weight of 0.556 for the lowest quartile and weights of 0.148 for the upper three quartiles. The weighted sum is then converted to a mean opinion score (MOS) using a nonlinear transformation consisting of the ratio of two logistic functions. Scores predicted using this approach were found to have an average correlation coefficient of 0.97 with subjective ratings of telephone speech processed through various encoding/decoding procedures, additive circuit noise, and quadratic and cubic distortion mechanisms (CCITT, 1987).

A. Coherence SII calculation

To compute the three-level CSI, the speech input signal envelope is divided into three amplitude regions. The calculation uses a block size of 16 ms with a Hamming window and 50% overlap between windowed segments. The magnitude of the signal in each segment is computed and stored over the duration of the sentence. The root-mean squared (RMS) level of the segments is then computed. The high-level segments are those at or above the overall RMS level. The mid-level segments are those between 0 and 10 dB below the overall RMS level, and the low-level segments are those between 10 and 30 dB below the overall RMS level.

The assignment of sentence envelope samples to the different levels is illustrated in Fig. 4 for the first sentence from HINT list 1. Approximately one-third of the segments fall into each of the three level regions. The high-level segments can be considered to be primarily vowels, the mid-level segments contain vowel-consonant transitions, and the low-level segments are mostly consonants and pauses. In terms of noise and distortion, the high-level segments will be most strongly affected by peak clipping, while the low-level segments will be most strongly affected by additive noise and center clipping.

For the computation, the output and input signals are first time-aligned. The high-level CSI (CSI_{High}) value is then computed by applying the coherence SII calculation procedure, including the auditory threshold, to just the high-level input segments and the corresponding output segments. The high-level MSC is computed using Eq. (6), but the summation is only for those segments classified as high-level input segments. The power spectra used in Eqs. (11)–(13) are similarly restricted to the high-level input segments and the corresponding output segments. The procedure is then repeated to give the mid-level CSI (CSI_{Mid}) and the low-level CSI (CSI_{Low}) values, after which the high-, mid-, and low-level CSI values are combined to give the prediction of speech intelligibility. In comparison with the CCITT (1986) procedure, the three-level CSI procedure is more computationally efficient since it does not require computing the signal magnitude histogram, and it requires spectrum and coherence calculations over three rather than four signal amplitude regions.

The correlations between the three level CSI values and the overall signal CSI and the subject results are presented in Table I. The correlation between the overall signal CSI and the average intelligibility scores is 0.722. The correlations between the low-level CSI and the scores and between the mid-level CSI and the scores are all higher than the correlations between the overall signal CSI and the scores. Thus the overall signal CSI is not the most useful statistic for describing the effects of the noise and distortion used in the experiment. Furthermore, the correlation between the low-level CSI and the overall signal CSI is only 0.703, indicating that the low-level CSI provides information about the noise and distortion that is absent in the overall signal CSI. Thus the combination of low-, mid-, and high-level

FIG. 4. Envelope for a sentence (HINT list 1, sentence 1) showing the level in dB relative to the RMS level. The dashed lines show the boundaries of the high, mid, and low level regions used for computing the CSI.
CSII values provides more information about the noise and distortion and should yield a more-accurate prediction of speech intelligibility and quality than provided by the overall signal CSII.

### B. Intelligibility model

The speech intelligibility model uses a linear weighting of the low-, mid-, and high-level CSII values. The weighted sum is then transformed using a logistic (or log-sigmoid) function to give the predicted proportion sentences correct (Gordon-Salant and Fitzgibbons, 1995). The weights are constrained to be positive numbers, and a bias term that adjusts the horizontal position of the curve is also included. The optimum fit of the model to the subject intelligibility scores was found using an unconstrained nonlinear minimization procedure (Nelder and Mead, 1965) to minimize the mean-squared error between the model and the average subject intelligibility scores. Let \( I_3 \) denote the intelligibility predicted by the model. The optimum solution is given by

\[
c = -3.47 + 1.84\text{CSII}_{\text{Low}} + 9.99\text{CSII}_{\text{Mid}} + 0.0\text{CSII}_{\text{High}},
\]

\[
I_3 = \frac{1}{1 + e^{-c}}.
\] (14)

The intelligibility appears to be dominated by the mid-level CSII, with some input from the low-level CSII. The weight for the high-level CSII is zero, so this term has no apparent effect on intelligibility in the context of this model. The high-level segments are predominantly vowel nuclei, which exhibit little variation over time in comparison with the high-level segments.

The three-level CSII intelligibility predictions provide a lateral shift of the prediction curve that minimizes the error between the prediction and the observed subject intelligibility scores.

The accuracy of the model predictions is shown in Fig. 5. The average intelligibility scores are plotted as a function of \( I_3 \) for additive noise, peak clipping, and center clipping. The model tends to underestimate the intelligibility for additive noise and to overestimate the intelligibility for center clipping. The overall accuracy of the model is high, however, with a correlation coefficient between the intelligibility scores and the model predictions of \( r = 0.94 \).

### VI. EXPERIMENT II: HEARING-IMPAIRED LISTENERS

The above results indicate that the three-level coherence SII procedure is accurate for normal-hearing listeners. Since an objective of this work is accurate prediction of intelligibility for hearing aids, the procedure must also be shown to work for hearing-impaired listeners. The intelligibility experiment was therefore repeated for a group of hearing-impaired subjects, and the results analyzed using the same three-level coherence SII technique that was used for the normal-hearing subjects.

#### A. Methods

Nine adult listeners with hearing loss of presumed cochlear origin (mean age = 49 years old; age range = 23–81 years) participated in experiment 2. (Audiometric thresholds of the listeners with hearing loss are plotted as part of Fig. 7.) Listeners underwent an audiometric evaluation during their initial visit. All of the listeners with hearing loss demonstrated test results that were consistent with cochlear impairment: normal tympanometry, the absence of excessive reflex decay, the absence of air-bone gap exceeding 10 dB at any frequency, and the absence of otoacoustic emissions in regions of threshold loss. All listeners were native English speakers.

As in experiment 1, the hearing-impaired listeners were tested individually in a double-walled sound-treated booth.

#### TABLE I. Correlations between the level CSII values, the overall signal CSII, and the average subject intelligibility scores.

<table>
<thead>
<tr>
<th></th>
<th>Low-level CSII</th>
<th>Mid-level CSII</th>
<th>High-level CSII</th>
<th>Overall CSII</th>
<th>Average Intelligibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-level</td>
<td>1</td>
<td>0.902</td>
<td>0.508</td>
<td>0.703</td>
<td>0.733</td>
</tr>
<tr>
<td>Mid-level</td>
<td></td>
<td>1</td>
<td>0.799</td>
<td>0.922</td>
<td>0.818</td>
</tr>
<tr>
<td>High-level</td>
<td></td>
<td></td>
<td>1</td>
<td>0.966</td>
<td>0.626</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>0.722</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
The stimuli were the same HINT sentence materials as were used for the normal-hearing subjects. The additive speech-shaped noise, peak-clipping distortion, and center-clipping distortion levels were set using the same procedures as for the normal-hearing listeners. The corrupted sentences were then equalized in level so that they had the same long-term 65-dB SPL amplitude as the unprocessed stimuli. The corrupted and level-adjusted sentences were then amplified using the NAL-R gain rule (Byrne and Dillon, 1986) computed for each hearing loss at the audiometric frequencies from 250 to 4000 Hz. The gains in dB were interpolated across frequency and used to design a 128-point linear-phase finite impulse response (FIR) filter through which the stimuli were processed. The sentences were presented monaurally to the listener’s test ear. The equipment and headphones were the same as for the normal-hearing subjects.

**B. Intelligibility scores**

The proportion of the sentences identified correctly for the additive noise, peak-clipping distortion, and center-clipping distortion is plotted in Fig. 6 for each of the nine hearing-impaired subjects. The general trend of the data is poorer performance as the amount of noise or distortion is increased, with the spread in the data also increasing as the noise or distortion increases. The speech recognition score for all of the subjects is 80% or better for no noise or distortion, and drops to close to zero for the worst-case noise and center clipping conditions. The overall trends in speech-recognition performance for the hearing-impaired listeners agree with those in the normal-hearing listener data presented in Fig. 1.

In general, the subjects with the greatest hearing loss have the poorest performance. For example, subject V (solid triangles) has a flat loss ranging from 30 dB at 250 Hz to an average of 50 dB above 2000 Hz and has nearly the best performance of the hearing-impaired subjects. Conversely, subject O (open circles) has a steeply sloping loss with over 70 dB of loss above 2000 Hz, and has uniformly poor performance.

**C. Hearing-impaired CSII**

The predicted intelligibility for the hearing-impaired subjects is given by the same $I_3$ calculation procedure as used for the normal-hearing subjects. The speech segments are each assigned to one of three level regions, and the coherence computed separately for the groups of low-, mid-, and high-level speech segments. The three coherence values are then combined using the offset and weights given in Eq. (14), followed by the logistic function transformation. The numerical coefficients used in Eq. (14) for the hearing-impaired listeners were exactly the same as those used for the normal-hearing listeners, and the upward spread of masking included in the SII calculation was as specified in the standard procedure for normal ears. The level distortion factor in the SII calculation was not used.

The only difference between the calculations for the normal-hearing and hearing-impaired subjects is that the calculation for the hearing-impaired listeners includes the hearing loss and the NAL-R amplification. These two adjustments affect the SDR as a function of frequency. The audiogram in the SII calculations is represented as an equivalent internal noise source, so the hearing loss reduces the SDR. The NAL-R amplification increases the signal gain in the regions of hearing loss and thus partly restores the SDR. The amplification, however, provides imperfect compensation; the SDR for all three signal levels is therefore reduced by the hearing loss, with the greatest effect occurring for the low-level CSII. The predicted speech intelligibility is there-
fore reduced both by the presence of noise and distortion in the signal and by the reduced audibility of speech sounds caused by the hearing loss.

The individual subject results are shown in Fig. 7, where the audiogram is presented for each subject along with the plot of the observed versus predicted intelligibility. The subjects are ordered by increasing severity of the high-frequency hearing loss. As explained above, the predictions for each hearing-impaired subject use the coefficients derived for the normal-hearing subjects in conjunction with the individual audiogram and NAL-R gain. The correlation coefficients between the predicted and observed intelligibility scores range from 0.88 (subject S) to 0.99 (subject P). There appears to be a slight trend of decreasing prediction accuracy with increasing hearing loss, but the $I_3$ predictions are quite accurate even for the most severe losses included in this study.

In Fig. 8, the average hearing-impaired intelligibility scores are plotted against the average $I_3$ predictions for each noise or distortion condition. The correlation coefficient between the average predictions and subject scores is 0.98. The
accuracy of the intelligibility predictions is high for all three forms of noise and distortion used in this experiment despite the widely different distortion mechanisms. Peak clipping predominately affects the high-level speech sounds, with reduced effects on the mid-level sounds, and noise and center clipping primarily affect the low-level speech sounds, again with reduced effects on the mid-level sounds. Because the $I_3$ calculation procedure divides the speech into high-, mid-, and low-level regions, it can effectively deal with the reduced audibility of speech information occurring at different speech amplitude levels for the noise and distortion used in this study.

VII. DISCUSSION

The experiments described in this paper deal with two issues. The first issue is extending the SII approach to include distortion as well as noise, and the second is predicting speech intelligibility for hearing-impaired listeners.

A. Distortion

Replacing the SNR in the SII calculation with the SDR computed from the coherence is an intuitively obvious extension of the principles underlying the SII. Authors of papers using coherence to measure distortion in hearing aids (Preves et al., 1989; Dyrlund, 1989; Preves, 1990; Gates, 1992; Fortune and Preves, 1992; Schneider and Jamieson, 1995; Gates, 2000) have made the assumption that there is a direct relationship between a reduction in coherence and a reduction in intelligibility. For any one type of noise or distortion, there will be a monotonic relationship between the increase in the amount of distortion and the reduction in the SDR. Thus earlier experiments that looked at just one type of distortion, such as peak clipping (Steeneken and Houtgast, 1980; Crain and van Tasell, 1994; Gates and Kozma-Spytek, 1994), can produce an accurate mapping of the SDR to intelligibility or sound quality for the distortion condition considered in the study. But as shown in this paper, those mappings may not be accurate if more than one type of distortion is considered. It is much more difficult to produce a relationship that is accurate for both center clipping, which most strongly affects the lower signal amplitude regions, and peak clipping, which most strongly affects the higher signal amplitude regions. The three-level coherence SII, on the other hand, combines the effects of the three different noise and distortion mechanisms considered in this paper into a single relationship that is accurate for all of them.

The three-level CSII approach divides the speech into low-, mid-, and high-amplitude regions. The three level CSII calculations provide much more information about the effects of the distortion on the speech than can be obtained by the SNR or the full-level CSII alone. In particular, the high-level CSII, which dominates the full-level CSII, has a weight of zero in the three-level calculation and thus no observed effect on intelligibility even though the high-level CSII is greatly reduced by peak clipping. Most of the speech information appears to be conveyed within a narrow 10-dB dynamic range from the RMS level of the sentence to 10 dB below, with a smaller contribution from the lower speech levels. The three-level CSII approach thus provides enough information to accurately model the reduction in intelligibility resulting from all three types of noise and distortion considered in this study, and does so in a manner consistent with observations on speech production and perception (Greenberg, 2005; Yoo et al., 2004). Increasing the number of levels to more than three may improve the accuracy of the intelligibility predictions, but it would also increase the number of degrees of freedom in the model and would thus increase the chances of over-fitting (fitting individual data points rather than the overall structure of the data) the limited amount of data available from the experiments.

B. Hearing-impaired listeners

Many researchers have tried to predict speech intelligibility in hearing-impaired listeners. The approaches typically taken in modeling the effects of hearing loss are the addition of a desensitization factor to the SII calculation (Pavlovic et al., 1986; Ching et al., 1998; Magnusson et al., 2001) or the inclusion of increased spread of masking (Ludvigsen, 1987; Rankovic, 1998) or individually measured masking functions (Hornsby and Ricketts, 2003). The desensitization factor is equivalent to a reduction in the SNR used in the SII calculation, with the amount of the reduction based on empirical observations of speech intelligibility in hearing-impaired listeners compared with those having normal hearing. Increasing the masking used in the SII calculation, either through increased spread of masking or the use of individual masking functions, has a similar effect in that it reduces the effective SNR.

The three-level CSII gives an alternative approach to modeling the effects of hearing loss. The CSII calculation for the hearing-impaired subjects requires no modification to the calculation procedure or masking functions; the audiogram and amplification used for the impaired ear are the only necessary inputs, after which the calculation for the hearing-impaired listeners proceeds exactly as for the normal-hearing listeners. The difference is that the three-level CSII incorporates a much more complete model of the effects of hearing loss on the audibility of speech. The low-level CSII is strongly affected by the hearing loss; the low-level speech sounds are much closer to threshold in the impaired than in the normal ear, and the low-level CSII is therefore much lower in the impaired ear even without noise or distortion. The addition of noise or distortion will reduce the low-level CSII even further, thus quickly removing the speech information provided in the low-level signal region. The mid-level CSII will experience similar effects, but with reduced severity given the higher signal level. The high-level CSII is affected the least, but the high-level CSII is not a factor in the $I_3$ calculation of Eq. (14).

These results indicate that audibility is an important factor in the difference in speech intelligibility between normal-hearing and hearing-impaired listeners. No modifications were made to the auditory filter shapes, upward spread of masking, or the importance functions used in computing $I_3$ for the hearing-impaired subjects, yet very accurate predictions of intelligibility were obtained over a wide range of hearing losses by incorporating just the audiogram. The SNR used in the standard SII calculation reflects the level of the
noise relative to the speech peaks and gives an imprecise indication of the effects of the noise on the low-level speech components. The desensitization factor or increased masking proposed previously for hearing-impaired listeners can thus be seen as attempts to take an imperfect measurement of speech audibility, the full-level SNR, and modify it to reflect the greater loss of audibility of low-level speech information experienced by the impaired ear.

The impaired auditory system exhibits processing deficits in addition to the elevated auditory threshold. These deficits include recruitment (Reger, 1936; Allen et al., 1990), broadened auditory filters (Tyler, 1986; Dubno and Schaefer, 1991), increased spread of masking (Gagné, 1983; Glasberg and Moore, 1986), and poorer temporal resolution (Oxenham and Bacon, 2003). It is possible that some of these effects are subsumed into the three-level coherence SII approach. For example, the loss of information in the low-level signal segments may be highly correlated with the audiogram, and thus predicted by the $I_1$ calculation, but could still be caused by poor temporal processing or other deficits in the impaired ear. The predictive value of the SII, or of the three-level coherence SII extension, is based on correlations and does not necessarily indicate a causal relationship. It is also possible that certain types of interference that were not considered in this paper, such as temporally fluctuating noise, may cause reductions in intelligibility for hearing-impaired listeners (Festen and Plomp, 1990; Peters et al., 1998) that are not predicted by the three-level coherence SII model.

C. Limitations and extensions

One limitation of the work reported in this paper is that the noise was broadband, and the distortion mechanisms were also applied to the broadband signals. These test conditions are appropriate for many communication systems. In hearing aids, however, frequency-dependent amplification is used, which can result in different amounts of noise or distortion in different frequency regions. Previous results (Kates and Kozma-Spytek, 1994; Kozma-Spytek et al., 1996; Tan et al., 2004) have shown that approaches based on the principle of coherence are accurate in predicting speech quality for frequency-dependent distortion conditions as well as for broadband distortion. However, experimental verification is needed to show that the three-level coherence SII approach is also valid for frequency-dependent distortion. Experiments are also needed to extend the range of noise and distortion conditions considered. Quantization noise can be caused by the digital processing in a hearing aid or communication system, and transducer problems such as displacement limiting and nonuniform driver magnetic fields can also cause distortion. An additional issue is the accuracy of the three-level coherence SII approach for fluctuating noise. Additional experiments are needed to validate and refine the metric for the wide range of noise and distortion conditions that occur in real life.

The three-level CSII approach should also lead to an improved procedure for predicting the quality of distorted speech. There is a strong correlation between intelligibility and sound quality (Preminger and Van Tasell, 1995; Eisenberg et al., 1998), which would lead one to expect a similarly high correlation between the intelligibility predicted by $I_4$ and sound quality for comparable noise and distortion conditions. While preliminary results are encouraging (Kates and Arehart, 2004), additional experiments are needed to extend the three-level coherence SII approach to produce an accurate speech quality metric for normal-hearing and hearing-impaired listeners.

VIII. CONCLUSIONS

Speech intelligibility is the primary concern in hearing aids and many communication systems. The SII has been developed and validated for additive noise and for bandwidth reduction, but the output of hearing aids is affected by distortion as well as by noise. The coherence between the input and the output of the device under test provides a tool for calculating the combined effects of noise and distortion in the processed output. It is therefore attractive to modify the SII calculation procedure to use the SDR computed from the coherence instead of the SNR, and thus include distortion along with noise.

Intelligibility was determined for broadband additive speech-shaped noise, peak clipping, and center clipping for normal-hearing and hearing-impaired subjects. A simple modification of the SII, using the SDR in place of the SNR but leaving the remainder of the procedure unchanged, was shown to be inadequate in predicting the intelligibility data for the normal-hearing listeners. The estimated intelligibility for peak clipping was similar to that for noise, but center clipping had much lower intelligibility for the same coherence SII values.

Much better accuracy was obtained by dividing the speech signal segments into three level regions. The high-level region consists of those segments at or above the sentence RMS level. The mid-level region ranges from the RMS sentence level to 10 dB below, and the low-level region ranges from RMS −10 dB to RMS −30 dB. The coherence SII is then computed separately for the segments in each level region. A linear combination of the CSII values, followed by a logistic function transformation, models the normal-hearing subject intelligibility scores with a correlation coefficient of 0.94.

The intelligibility experiment was then repeated for a group of hearing-impaired listeners. The distorted speech was amplified using the NAL-R gains computed for each hearing loss, and the three-level CSII calculation was adjusted for the hearing loss by including the audiogram and the NAL-R gain in the constituent CSII calculations. The intelligibility predictions were accurate across the entire range of hearing losses used in the experiment, and the average across subjects for the different noise and distortion conditions had a correlation coefficient of 0.98. The three-level CSII thus appears to be an accurate procedure for estimating intelligibility for the conditions of distortion and noise considered in this paper, and it is as effective for hearing-impaired listeners as it is for those with normal hearing.
ACKNOWLEDGMENTS

Author KKH was supported by the Whitaker Foundation. The information provided in this paper was also supported in part by Grant/Cooperative Agreement Number UR3/CCU824219 from the Centers for Disease Control and Prevention (CDC). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of CDC. The authors thank Rosalinda Baca, Denise Kidd, and Jessica Rossi-Katz for assistance in the data collection, and Amit Das for development of the software used for the listener tests. The authors also thank Professor Brian Moore for helpful comments on earlier drafts of this paper.


